MACHINE LEARNING IN BUSINESS

MIS710 – Assignment 1

Road Safety Analysis: Predicting and Addressing Blackspot Incidences in Victoria

Subject Code: MIS710

Subject Name: Machine Learning in Business

Class Group: Wednesday (15:00 – 16:30)

Tutor: Dat Le

Student Name and ID: Cam Ha Nguyen – s223546667

Word Count: 2349

DEAKIN
BUSINESS
SCHOOL

Table of Contents

Executive Summary

The purpose of this analysis is to address the rising concern of road accidents and fatalities on Victorian roads by identifying blackspots and their contributing factors. VicCrashAnalytics collaborates with stakeholders to enhance road safety, allocate interventions efficiently, and shape data-driven policies. The analysis involves a comprehensive approach, utilizing data cleansing, Exploratory Data Analysis (EDA), and machine learning techniques.

Through meticulous data cleansing, the blackspot dataset's integrity is ensured. EDA offers insights into road segment characteristics and their association with blackspots. Logistic Regression is utilized as the predictive model due to its interpretability and efficiency. The model undergoes iterative refinement to enhance its predictive performance. The resulting model exhibits consistent accuracy, and the ROC curve and AUC values confirm its effectiveness in blackspot prediction.

The analysis uncovers actionable insights. Locations with amenities like supermarkets and schools are linked to higher blackspot likelihoods, suggesting the need for tailored traffic management. Intersections and segments with traffic lights exhibit higher blackspot occurrences, warranting targeted safety measures. The presence of liquor license venues also correlates with blackspots, necessitating collaborative efforts to address traffic-related concerns.

In conclusion, VicCrashAnalytics' analysis blends data insights with collaborative approaches to mitigate blackspot incidents on Victorian roads. By effectively identifying and addressing risk-prone areas, road safety can be significantly improved, leading to lowered accident rates, optimized interventions and well-informed policies, thereby fostering a safer road environment.

1. Business Understandings and Problem Statement

In recent years, road accidents have become a global issue, ranking as the ninth leading cause of mortality worldwide (Jadhav et al., 2020). This challenge results in innumerable casualties, injuries, and fatalities each year, as well as enormous economic losses (Santos et al., 2021). The number of fatalities on Australian roadways increasing by more than 6% in the past year, exceeding long-term safety goals and remaining much higher than before the Covid epidemic (Visontay, 2022). Specifically, the pressing issue at hand involves the prevalence of blackspots on Victorian roads, where there has been a startling 9.7% increase in road deaths (Visontay, 2022). In order to successfully address this concerning pattern, it is essential to conduct a thorough and comprehensive analysis. By employing supervised machine learning techniques on the provided blackspot dataset, VicCrashAnalytics seeks to predict the likelihood of blackspots and uncover the underlying factors contributing to accidents. This endeavor involves a collaborative effort with stakeholders including the Victorian Department of Transport (DOT), law enforcement agencies, local communities, and road users. The anticipated outcomes include a substantial enhancement in road safety, a reduction in accident rates, informed resource allocation for interventions, and a data-informed policy landscape. This effort perfectly aligns with the DOT's overarching commitment to elevating road safety and is consistent with broader road safety initiatives, underpinning the need for a proactive and data-driven approach.

2. Data Understanding, Cleansing, EDA and Insights

2.1. Data Understanding

The dataset consists of 5326 rows representing observations and 36 columns serving as variables, with variable type including both numerical and categorical values. The dataset also provides diverse information about road segments and their surrounding characteristics. Each entry is characterized by a unique identifier and several attributes that provide insights into road safety conditions. The list of independent variables includes the full name and type of road segment, demographic breakdown by age groups, language spoken at home, family composition, household car ownership, dwelling types, occupation distribution, socioeconomic index, land use purposes, number of liquor licensed venues, amenities availability, which will be evaluated to find potentialt blackspot predictors. On the other hand, the last column, “blackspot”, functions as the dependent variable and target for the Machine Learning model.

2.2. Data Cleansing and Preparation

In the pursuit of accurate and reliable analysis, the dataset underwent a careful process of data cleansing and preparation. This crucial step ensures that subsequent insights are based on a solid foundation of good-quality information. To enhance the readability and interpretability of the dataset, column names were updated with intuitive descriptors .

Following the initial step taken to refine column names, the next step involved preliminary data transformation to make the dataset more understandable. In particular, information within columns such as ''has_supermarket'', ''has_primary_school'', ''has_secondary_school'', ''has_speed_sign'', and ''has_traffic_signal'' underwent conversion. The numeric values of "1" and "0" were replaced with categorical labels “Yes” and “No”. This adjustment aids in smoother data processing in later stages.

Missing data, a common concern in any dataset, was attentively addressed. Two columns, specifically "age_65_plus" and "liquor_license_venues," contained 9 and 6 missing values respectively, spread across 11 rows. Given the type of missing data does not follow a Not Missing At Random (NMAR) pattern, it is recommended to set a threshold 5% of the total entries for acceptable missing data. This implies that a maximum of 266 rows in the blackspot dataset could be discarded or imputed, as suggested by Bennett (1999) and Cheema (2014). Hence, 11 rows (0.2% of the total observations) were eliminated from the dataset. This choice ensures that the dataset's integrity remains intact while minimizing the impact of data loss.

Further examination of the dataset shows no record for duplicate entries, indicating the reliability and effectiveness of the collection process.

In the subsequent step, outliers and anomalous values were addressed. Numerical columns underwent investigation for potential outliers using the Interquartile Range (IQR) method, highlighting values significantly deviating from expected ranges. While a considerable number of rows (2708 out of 5311) displayed outlier-like behavior, no removal was undertaken given the dataset's focus on road segment characteristics. Eliminating all these rows would result in substantial data loss and potentially compromise the dataset's overall representation and analysis. Nonetheless, specific columns like “couples_with_children” and “white_collar_occupation” had instances where distribution percentages exceeded 100%.

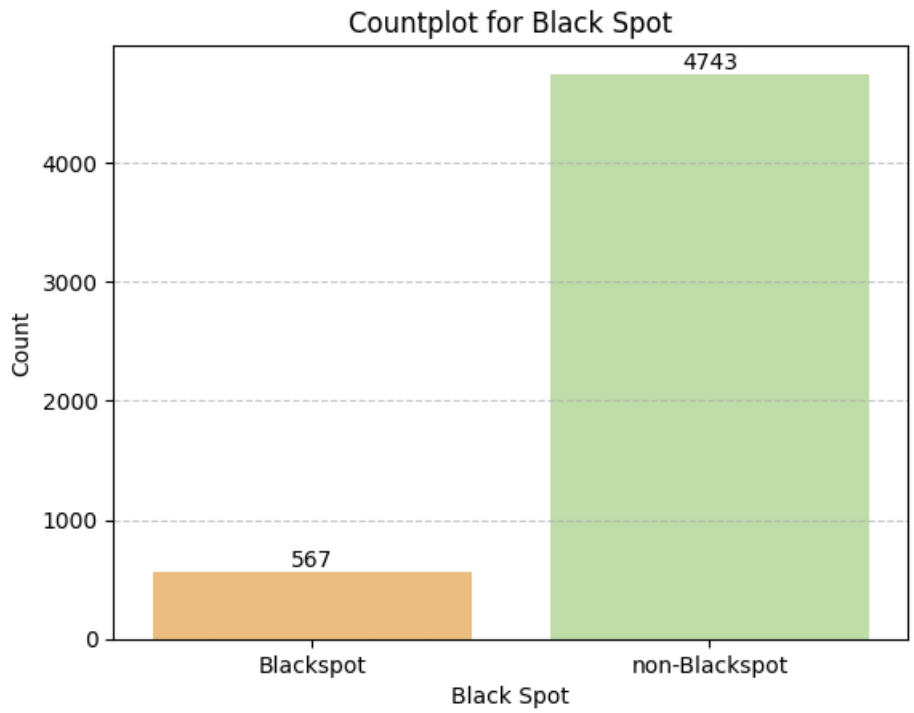
These abnormal values were systematically removed to maintain consistency within the data's patterns.

After cleaning, 5310 rows and 36 columns remain, forming a refined dataset ready for analysis and model development.

2.3. Exploratory Data Analysis and Insights Gained

The EDA process delved into the dataset for key road safety and blackspot insights. Through a combination of visualizations and data relationships, a comprehensive understanding of the dataset's characteristics and their impact on blackspots was developed.

The dataset encompasses 567 blackspots and 4743 non-blackspot road segments, revealing a clear imbalance in class distribution as blackspot instances only account for 10.68% (Figure 1). This imbalance is essential to acknowledge, as it can impact model performance.




```
Counts of Blackspots and Non-Blackspots:
non-Blackspot    4743
Blackspot         567
Name: blackspot, dtype: int64

Percentage of Blackspots and Non-Blackspots:
Percentage of Blackspots: 10.68%
Percentage of Non-Blackspots: 89.32%
```

Figure 1 - Distribution of blackspot and non-blackspot

As road type is associated with accidents (Santos et al., 2021), analyzing blackspots distribution across different road types reveals that road and street segments as most common, with 477 out of the total 567 blackspots (Figure 2).

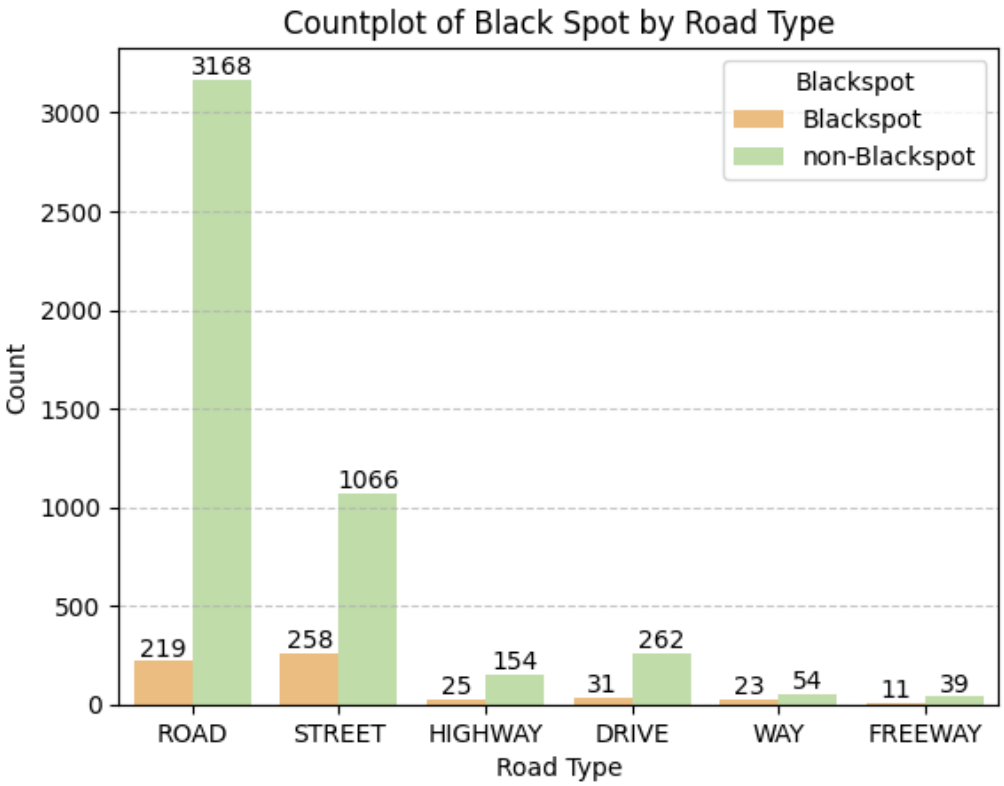


Figure 2 - Distribution of blackspot by road type

Age distribution within road segments varies across age groups (Figure 3). Notably, the age groups 25-44, 45-64 and 65+ have higher medians compared to other groups, implying potential blackspot association (Figure 4). For instance, a higher percentage of middle-aged drivers (25-44 years) could suggest a greater commuter population, potentially leading to specific types of accidents during peak traffic hours.

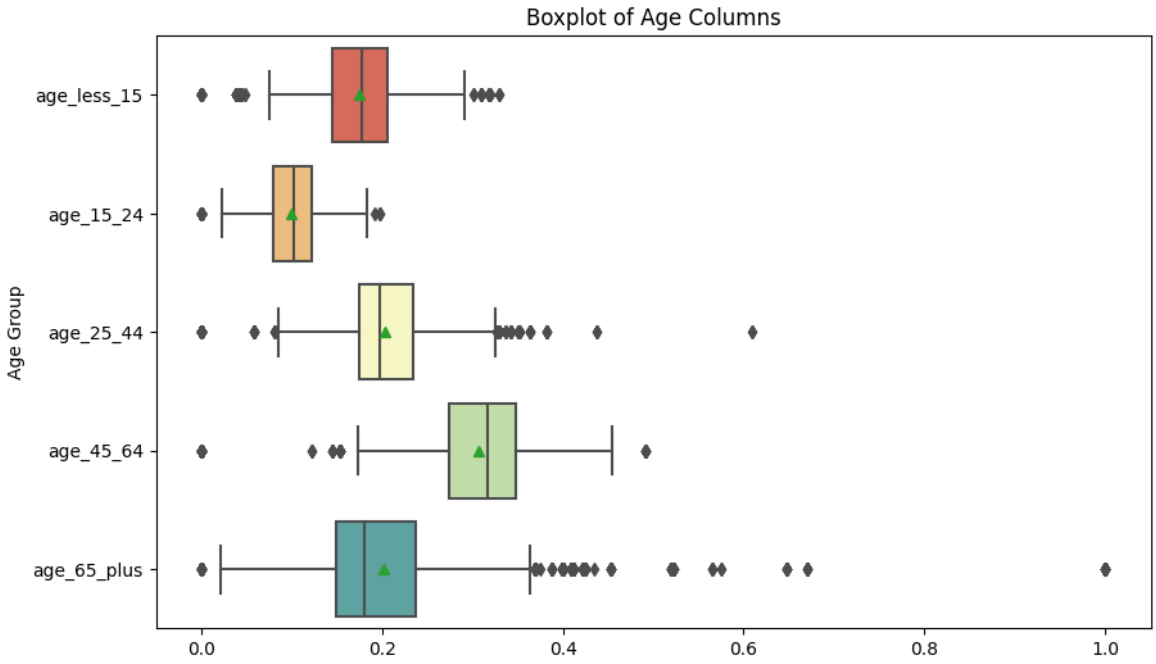


Figure 4 - Box plot for age group percentage distribution

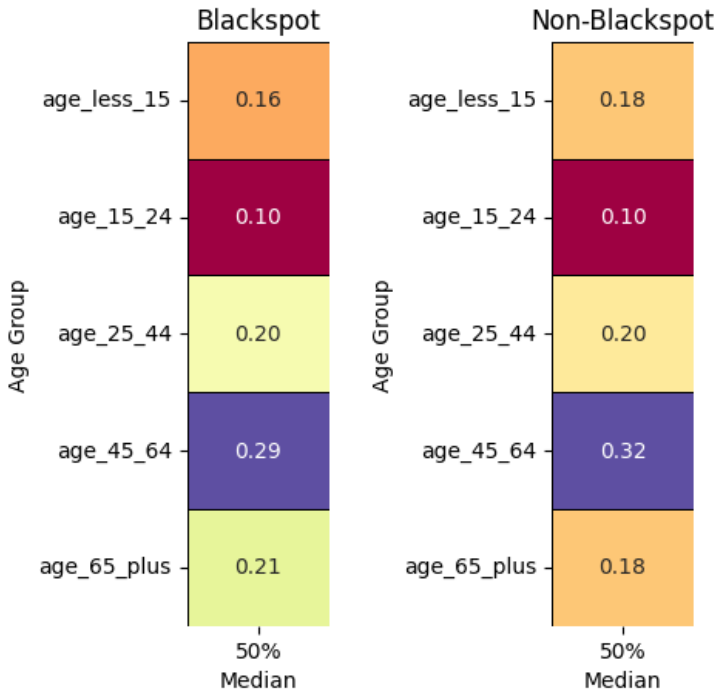


Figure 5 - Median comparison among different age groups

Exploring the relationship between occupation and blackspot occurrences provides insightful findings. While Borg and Compton (2015) link blue-collar workers to drowsiness-related accidents, a comparison of median occupation percentages between blackspot and non-

blackspot cases indicates no substantial difference (Figure 3). This suggests that occupation might not be a dominant contributor to blackspot formation.

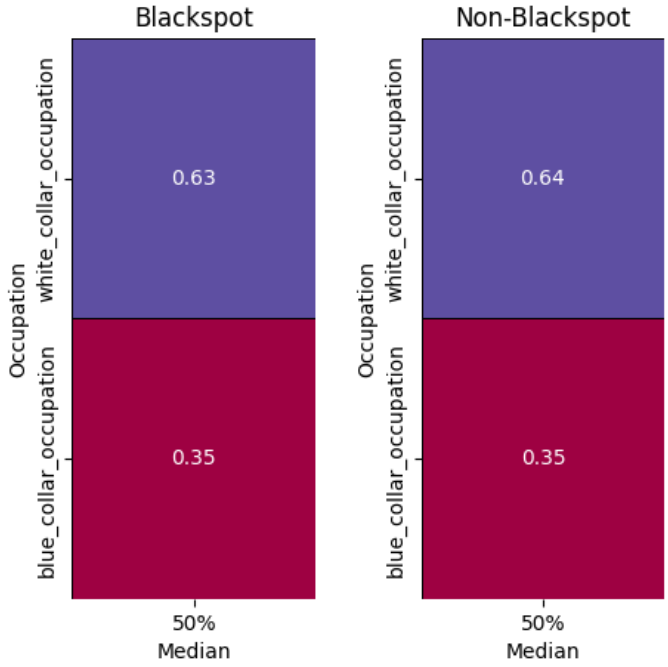


Figure 6 - Median comparison among different occupations

Figure 7 shows more blackspots in intersection road segments than non-intersection ones. Additionally, non-intersection road segments exhibit a notably lower occurrence of non-blackspots in comparison to blackspots. The discrepancies highlight the significance of intersections as potential accident-prone zones

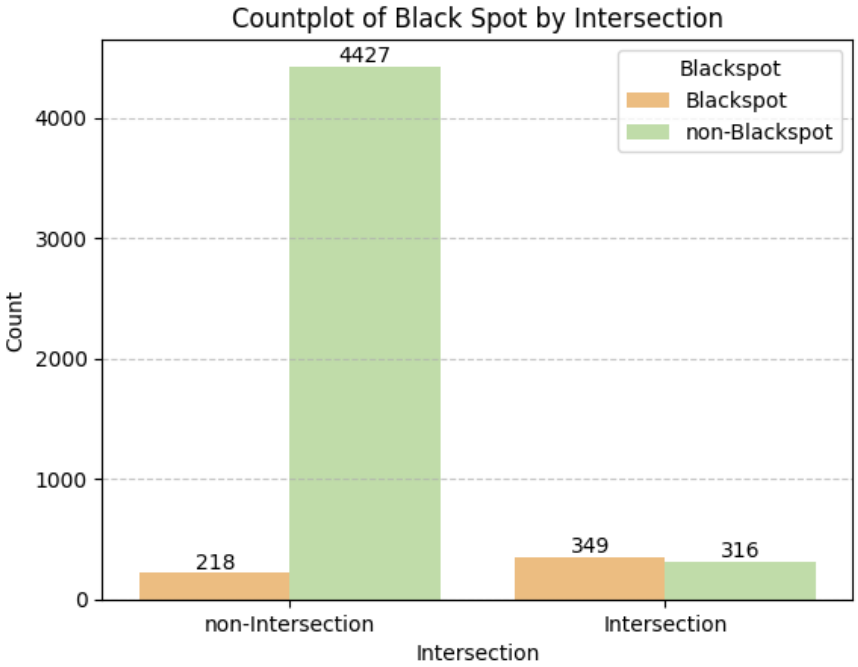


Figure 7 - Countplot of blackspot by intersection

The presence of amenities like supermarkets, primary schools, and secondary schools on road segments tends to attract higher traffic flow from diverse road users, which can lead to congestion and complex traffic conditions. As depicted in Figure 8, road segments hosting these amenities tend to have a higher ratio of blackspots.

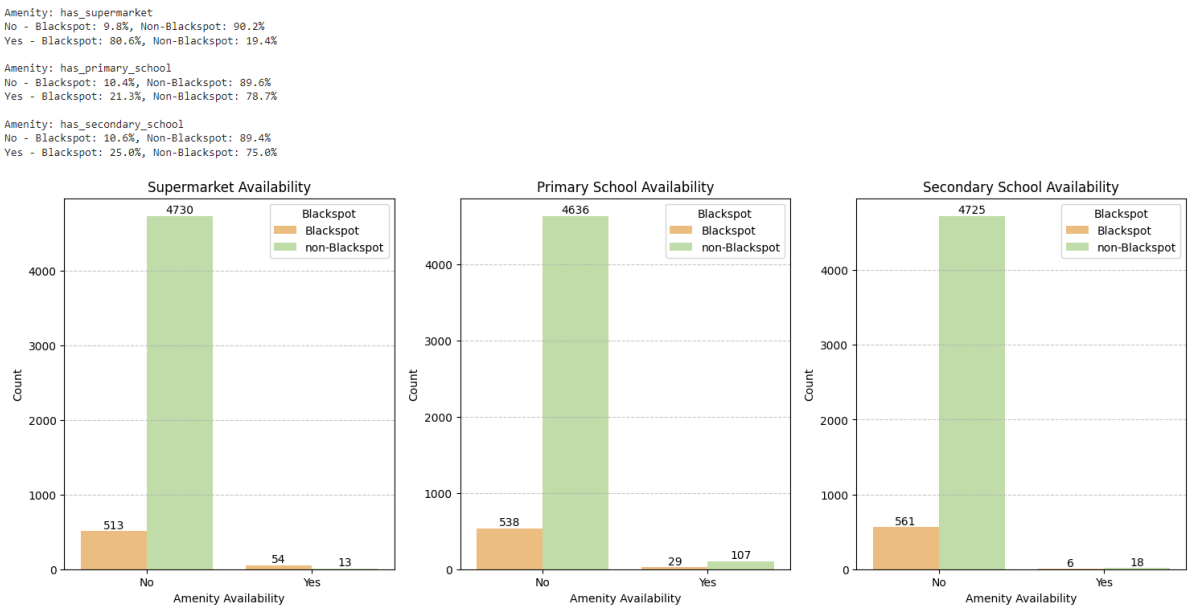
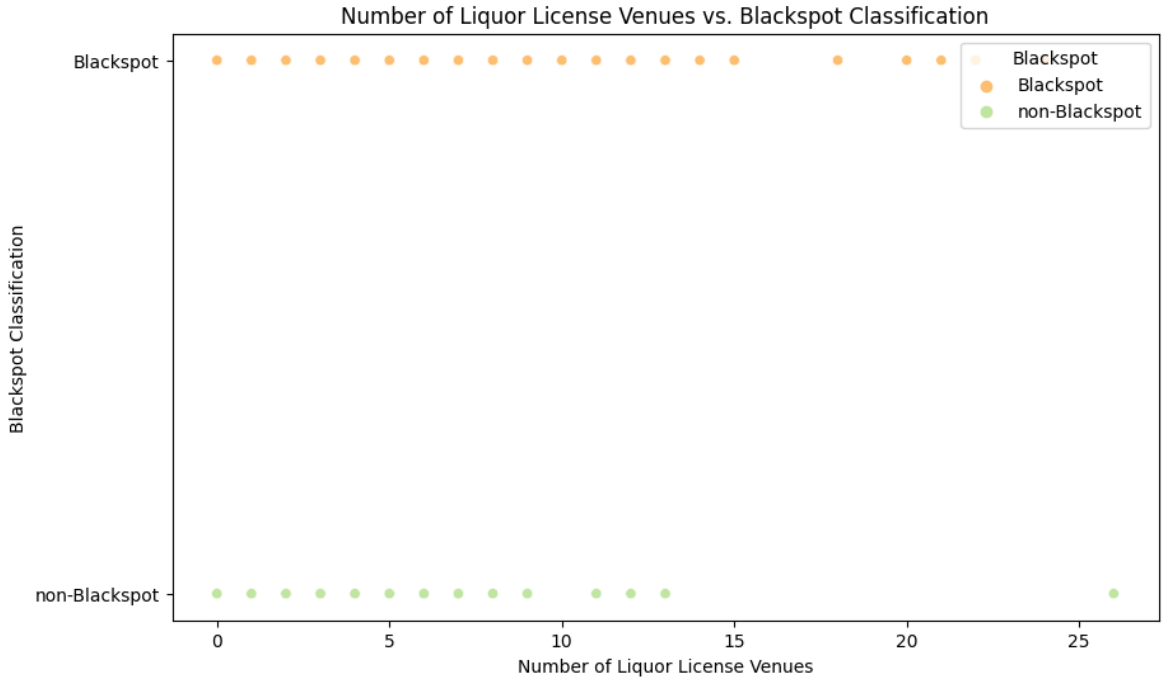


Figure 8 - Blackspot ratio comparison by amenities availability

Out of 567 blackspot cases, 235 cases (41.5%) have at least one liquor license venue, while among 4743 non-blackspot cases, 447 (9.4%) have such venues (Figure 8). The percentage of blackspot cases with liquor license venues is significantly higher compared to non-blackspot cases, suggesting potential association between these two variables.




```
Number of Blackspot cases with Liquor License Venues > 0: 235
Total Number of Blackspot cases: 567
Percentage of Blackspot cases with Liquor License Venues > 0: 41.4%

Number of Non-Blackspot cases with Liquor License Venues > 0: 447
Total Number of Non-Blackspot cases: 4743
Percentage of Non-Blackspot cases with Liquor License Venues > 0: 9.4%
```

Figure 9 - Comparison of liquor license venues in blackspot and non-blackspot cases

As the exploration into the presence of traffic lights unfolds, an interesting observation emerges. Road segments equipped with traffic signals exhibit a greater frequency of blackspots (Figure 10). This observation leads to speculation about the possibility of faulty traffic lights contributing to this trend, prompting a need to closely analyze the dynamics that contribute to this phenomenon.

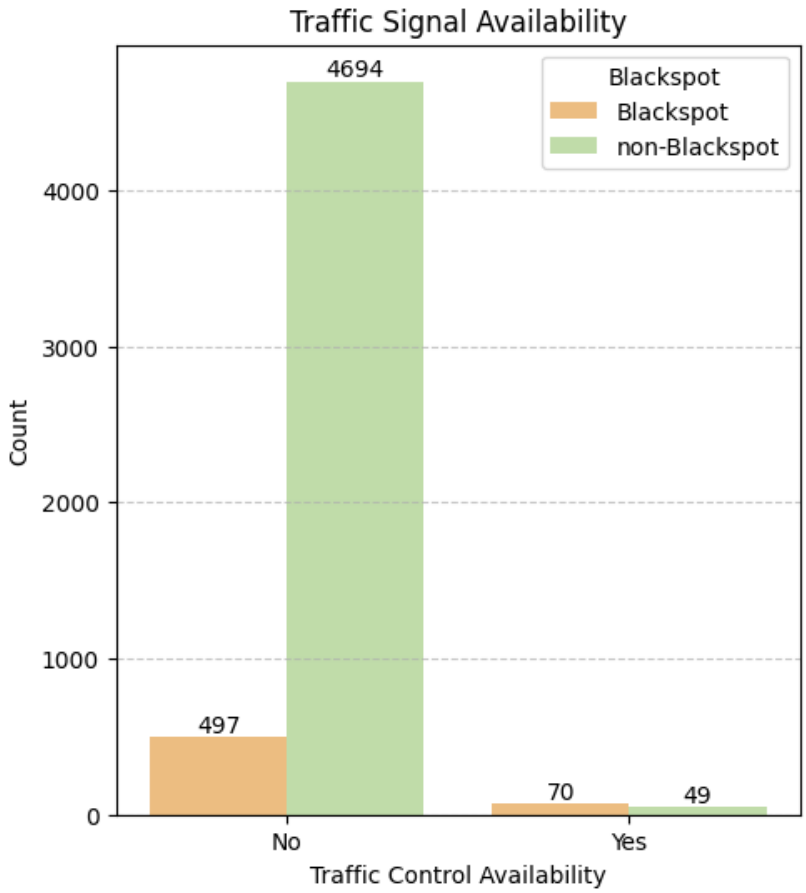


Figure 10 - Countplot of blackspot by traffic signal availability

Transitioning to multivariate analysis, the report delves into exploring dataset relationships through a correlation matrix. To facilitate this, strategic encoding steps were taken to convert categorical data to numeric format.

Categorical variables related to conveniences and infrastructure like "has_supermarket", "has_primary_school", "has_secondary_school", "has_speed_sign", and "has_traffic_signal" were converted to binary representation. In addition, "road_type" was transformed into dummy variables for multivariate analysis, resulting in six new columns for each road type. Irrelevant columns like "id", "age_18_plus", “full_road_name”, and “road_name” were excluded.

With data encoded and pruned, a correlation matrix was calculated to present the relationships between numeric variables (Figure 11). Colors represented relationship strength and direction, with cooler blue for negative and warmer red for positive correlations. Annotations provided correlation coefficients from -1 to 1, with 0 denoting no correlation.

This visual exploration uncovers the intricate interplay of variables, guiding feature selection by identifying influential factors for predicting blackspots.

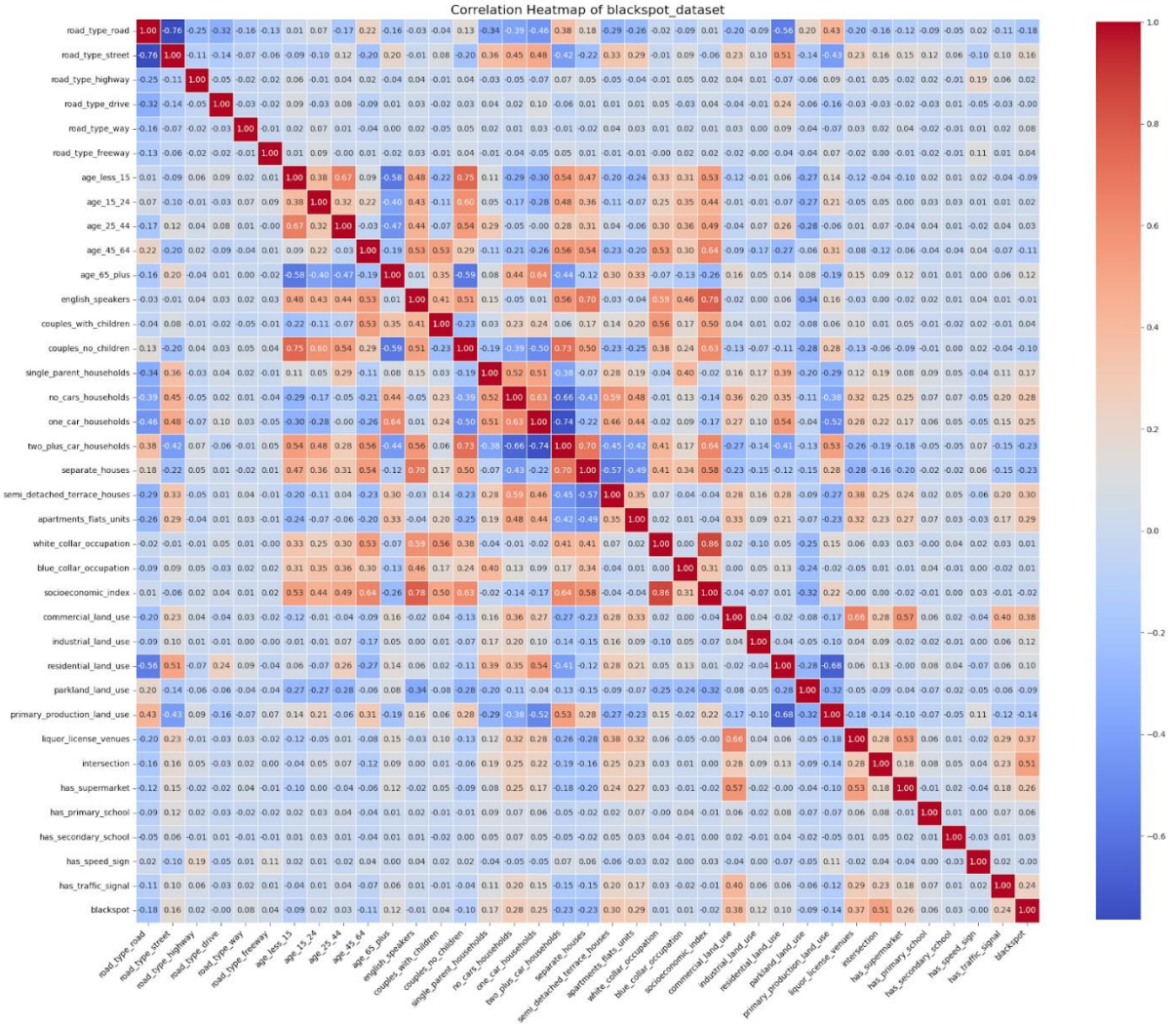


Figure 11 - Correlation matrix of blackspot dataset

3. Machine Learning Approach

The chosen model for this classification task is Logistic Regression. This selection was informed by the nature of the problem at hand – predicting the binary outcome of blackspot presence denoted as "1", or absence denoted as "0".

Three model iterations were executed to train and fine-tune the model, with variations in feature selection strategies. The initial iteration involved selecting features with absolute correlation values in the top 25% range of all coefficients. The criterion considered both

positive and negative coefficients, as long as they indicated stronger correlation (closer to 1) with the target variable "blackspot" (Figure 12).

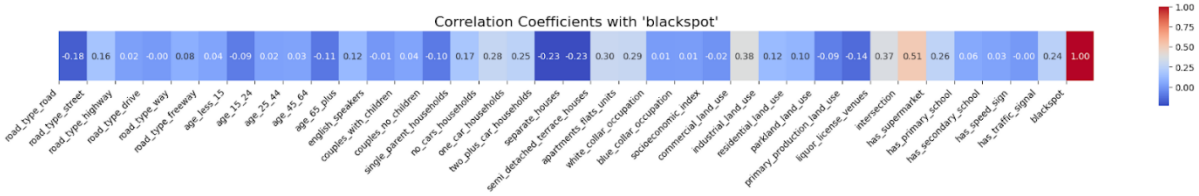


Figure 12 - Corelation coefficients between all features and blackspot

Subsequently, in the second and third iterations, a more rigorous feature selection method was applied (Appendix B). We employed backward elimination utilizing the statsmodels library in Python (Figure 13). This process entailed iteratively removing the feature with the least statistic significance, determined by its p-value. A p-value below 0.05 is typically deemed statistically significant.

```
Optimization terminated successfully.
         Current function value: 0.222588
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:              blackspot   No. Observations:                 4248
Model:                          Logit   Df Residuals:                     4238
Method:                           MLE   Df Model:                            9
Date:                Thu, 10 Aug 2023   Pseudo R-squ.:                  0.3462
Time:                        23:56:57   Log-Likelihood:                -945.55
converged:                       True   LL-Null:                       -1446.1
Covariance Type:            nonrobust   LLR p-value:                9.623e-210
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           -3.6265      0.168    -21.530      0.000      -3.957      -3.296
no_cars_households              -0.2213      2.207     -0.100      0.920      -4.547       4.104
one_car_households               1.2729      0.647      1.968      0.049       0.006       2.540
semi_detached_terrace_houses     2.5555      0.859      2.975      0.003       0.872       4.239
apartments_flats_units           2.5888      1.103      2.346      0.019       0.426       4.751
commercial_land_use              2.2935      0.637      3.598      0.000       1.044       3.543
liquor_license_venues            0.1670      0.043      3.870      0.000       0.082       0.252
intersection                     2.6427      0.128     20.633      0.000       2.392       2.894
has_supermarket                  0.2258      0.548      0.412      0.680      -0.847       1.299
has_traffic_signal               0.5351      0.300      1.783      0.075      -0.053       1.123
================================================================================================
```

Figure 13 - Logit regression results from statsmodel

In each iteration, the dataset was split into training and testing subsets, with 80% assigned to training and 20% to testing. This enabled the model to be trained, tested, and evaluated using the updated feature sets. Reproducibility was ensured through a consistent random seed.

StratifiedKFold cross-validation was used for validation, ensuring that the distribution of the target variable was preserved across folds (Prusty et al., 2022). The last iteration of logistic regression model results in seven predictors selected for blackspot classification (Apendix C).

4. Model and Performance Metrics

In all three iterations, the confusion matrix displayed consistency in the count of true positives, true negatives, false positives, and false negatives, indicating that the model's classification results remained largely unchanged (Figure 14-16).

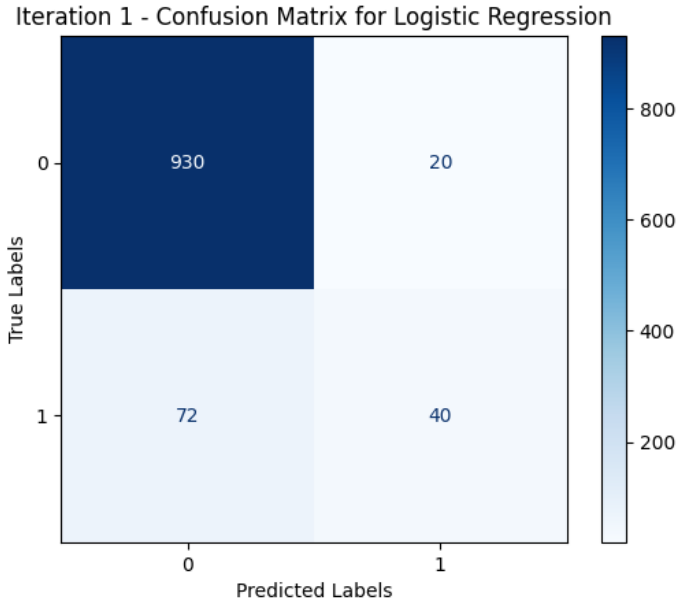


Figure 14 - Confusion matrix in first iteration

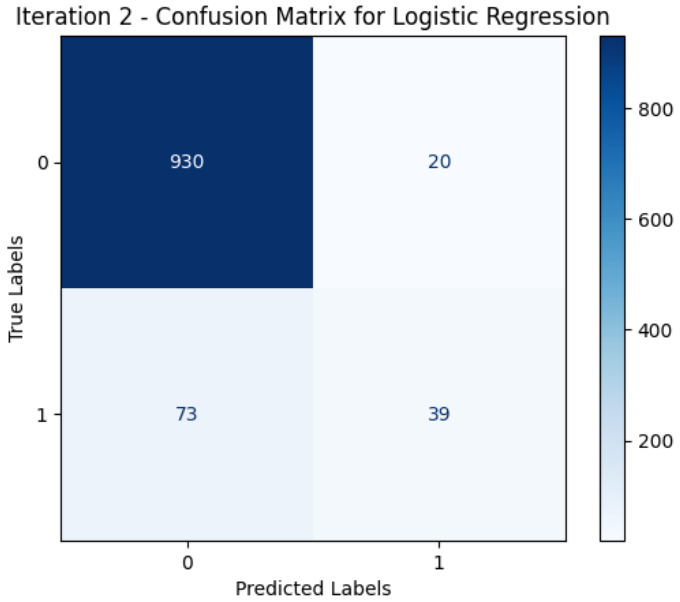


Figure 15 - Confusion matrix in second iteration

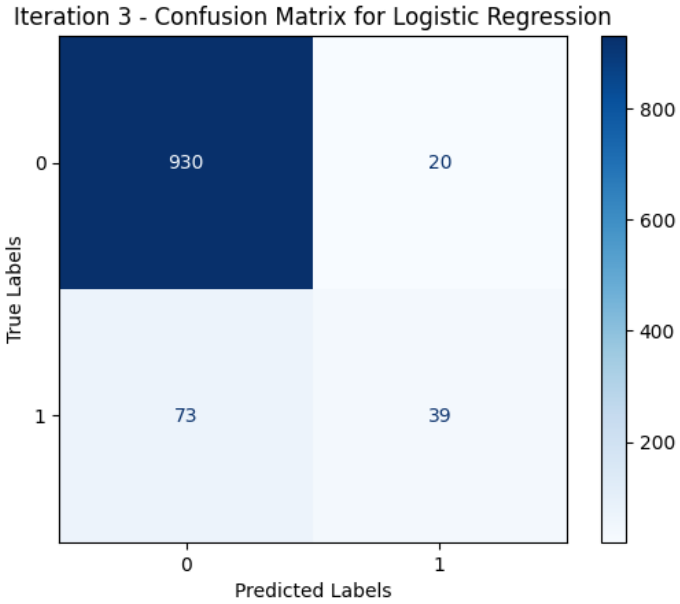


Figure 16 - Confusion matrix in third iteration

Commonalities emerged across all three iterations in terms of accuracy, precision, recall, and F1 score. The model's accuracy consistently hovered around 0.918 with a minor deviation of +/- 0.007, signifying that the model achieved accurate classifications for approximately 91.8% of instances and showcasing the model's strong ability to generalize to new data (Figure 17). This stability is also evident in the classification reports, where metrics display minimal variability (Figure 18). However, there is room for improvement in terms of both precision

and recall, particularly for the positive class (blackspot), to enhance the model's performance in identifying positive instances. In the final iteration, the model's precision is around 72.6%, with a variation of about ±7.4%. This means that when the model predicts an instance as a blackspot, it is correct 72.6% of the time. The relatively lower precision suggests that there might be false positives. The explanation for this lies in the class imbalance as the dataset contains 4743 (89.3%) non-blackspot labels but only 567 (10.7%) blackspot ones, leading to higher precision for the dominant class but lower precision for the minority class. The recall remains consistent at 37.4%, indicating that the model is still missing a significant portion of actual positive instances.

	Iteration 1	Iteration 2	Iteration 3
Accuracy:	0.918 +/- 0.007	0.918 +/- 0.007	0.918 +/- 0.007
Precision:	0.726 +/- 0.074	0.725 +/- 0.071	0.726 +/- 0.074
Recall:	0.374 +/- 0.061	0.374 +/- 0.061	0.374 +/- 0.061
F1 Score:	0.490 +/- 0.058	0.490 +/- 0.059	0.490 +/- 0.058

Figure 17 - Cross-validation metrics for logistic regression in three iterations

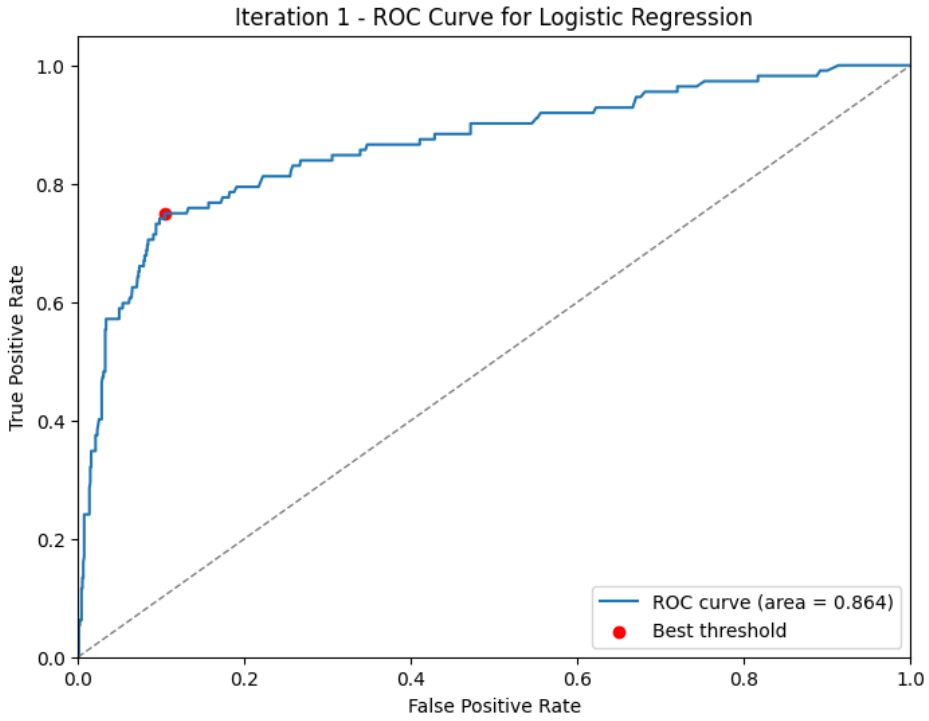
ITERATION 1				
	precision	recall	f1-score	support
0	0.93	0.98	0.95	950
1	0.67	0.36	0.47	112
accuracy			0.91	1062
macro avg	0.80	0.67	0.71	1062
weighted avg	0.90	0.91	0.90	1062

ITERATION 2				
	precision	recall	f1-score	support
0	0.93	0.98	0.95	950
1	0.67	0.35	0.46	112
accuracy			0.91	1062
macro avg	0.79	0.66	0.70	1062
weighted avg	0.90	0.91	0.90	1062

ITERATION 3				
	precision	recall	f1-score	support
0	0.93	0.98	0.95	950
1	0.66	0.35	0.46	112
accuracy			0.91	1062
macro avg	0.79	0.66	0.70	1062
weighted avg	0.90	0.91	0.90	1062

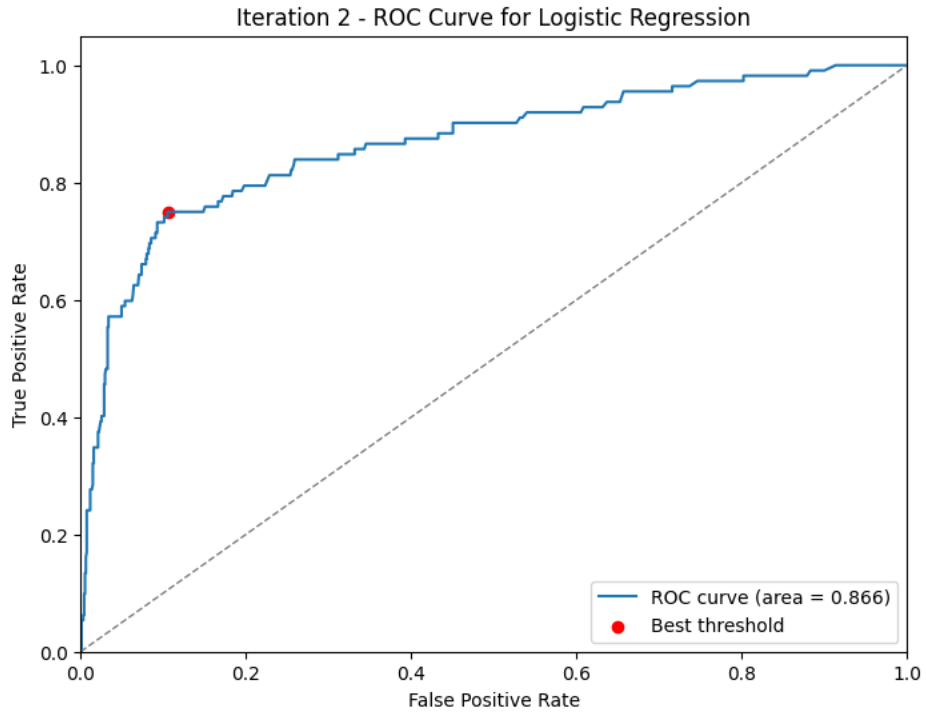
Figure 18 - Classification metrics for logistic regression in three iterations

The ROC curve and AUC evaluate binary classification performance across different thresholds, showing trade-offs between true positive rates (TPR) and false positive rates (FPR) (Oswald, 2020). ROC curve plots TPR against FPR at various thresholds, while AUC assesses the collective performance of ROC curve by quantifying the area beneath it (Lee, 2019). Throughout three iterations, AUC remained high with the last two iterations exhibiting slightly elevation at 0.866 compared to 0.864 in iteration 1(Figure 19-21). The best thresholds being consistently in the range of 0.074 implies relatively conservative positive classifications, which could be due to nature of the problem or data distribution.




```
Best threshold: 0.07418995850391812
```

Figure 19 - ROC, AUC and best threshold in first iteration




```
Best threshold: 0.07263734836040972
```

Figure 20 - ROC, AUC and best threshold in second iteration

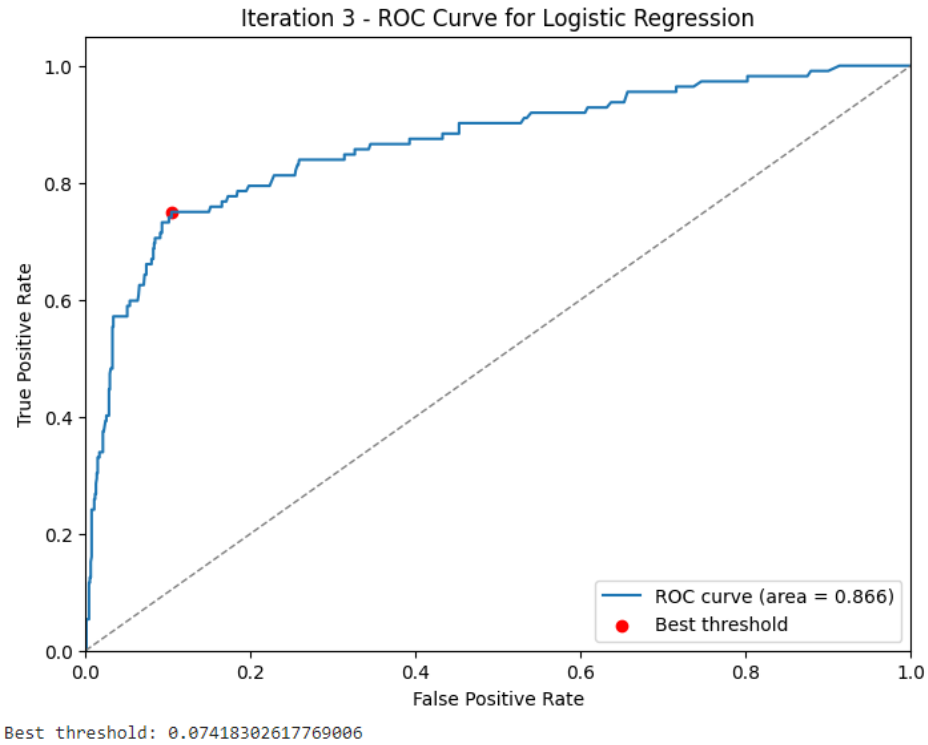


Figure 21 - ROC, AUC and best threshold in third iteration

5. Pros and Cons of the Model

Regarding the logistic model's strength, its low complexity makes implementation and interpretation easier. It not only gauges the appropriateness of predictors (coefficient value) but also their association direction (positive or negative) (Lee, 2019). The model's tendency to avoid overfitting (Dreiseitl and Ohno-Machado, 2022) is also an advantage. Additionally, given the performance in the blackspot case, the model is particularly adept at classifying non-blackspot instances, demonstrated by consistently high precision and recall values for class "0". This strength aligns well with the practical objective of effectively identifying non-blackspot road segments to prioritize intervention efforts.

However, certain limitations must be acknowledged. Logistic regression's major limitation is the assumption of linearity between the dependent and independent variables (Pant, 2022). Next, presence of imbalanced data can affect the model's ability to predict the minority class effectively (Zhang et al., 2021), as seen in blackspot instances. In addition, multicollinearity, where predictor variables in a logistic regression model are highly correlated, was not considered during model construction. This phenomenon can cause unstable estimates and

inaccurate variances, affecting confidence intervals and hypothesis tests (Midi et al., 2013). A range of solutions is available for logistic regression, including increasing sample size, ignoring one of the correlated variables, and combining variables into an index (Senaviratna, 2019). Without increasing the sample size, omitting one of the correlated variables can significantly reduce multicollinearity (Midi et al., 2013).

Given the classification challenge and the potential for improved performance in business practice, it is advisable to compare four or five algorithms and select the best Machine Learning model to implement in production (Verdhan, 2020).

6. Business Solutions and Recommendations

Utilizing analysis from EDA and last iteration of the logistic model above, several actionable insights are considered for road safety intervention.

Our EDA reveals that the availability of supermarkets, primary and secondary schools is linked to a higher likelihood of blackspot occurrence. Similarly, features like dwelling types (apartments, terrace houses) and commercial land use selected for the model show a strong positive correlation with blackspots. These elements seem to contribute to densely populated road segments, affecting traffic flow and leading to blackspots. To tackle this issue, potential solutions include optimizing traffic management, implementing speed reduction measures near these areas, and enhancing pedestrian safety measures.

As road and street segments are most closely associated with blackspots, there is a need for tailored safety measures for these road types, potentially including enhanced signage, speed control measures, and better lighting.

Regarding intersection vulnerability, vehicle detection technology such as in-road sensors and radar technology can be implemented to improve safety (Victorian Government, 2023).

The elevated percentage of blackspot cases involving liquor license venues underscores a possible correlation. Collaborative endeavors with local authorities to address traffic-related issues linked to these venues could result in safer road segments.

Additionally, it iss crucial to inspect road segments with traffic signals, as there is a potential for faulty signals to contribute to road accidents.

While the logistic regression model provides predictive insights, road segments as blackspots have significant implications for the model's effectiveness and the overarching safety goals. The criteria for categorizing a road segment as a blackspot should be defined collaboratively by stakeholders, including traffic authorities, urban planners, and safety experts. These experts contribute domain-specific knowledge that can refine the model's features, enhance the accuracy of predictions, and ensure that the chosen criteria align with real-world safety concerns.

References

Borg, K., & Compton, S. (2015). *Road Safety Monitor 2015*. The Social Research Centre. https://www.tac.vic.gov.au/__data/assets/pdf_file/0010/180883/TAC-RSM-2015-Main-Report-FINAL-05012016.pdf

Campesato, O. (2020). *Python 3 for Machine Learning*. Mercury Learning & Information.

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research*, *84*(4), 487-508. https://doi.org/10.3102/0034654314532697

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, 35(5-6), 352-359. https://doi.org/10.1016/S1532-0464(03)00034-0

Jadhav, A., Jadhav, S., Jalke, A., & Suryavanshi, K. (2020). Road accident analysis and prediction of accident severity using machine learning. *International Research Journal of Engineering and Technology (IRJET)*, *7*(12).

Lee, W. (2019). *Python machine learning*. John Wiley & Sons, Incorporated. https://ebookcentral.proquest.com/lib/deakin/detail.action?docID=5747364#

Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, 13(3), 253-267. https://doi.org/10.1080/09720502.2010.10700699

Pant, M. (2022). Advantages and Disadvantages of Logistic Regression. *Kaggle*. https://www.kaggle.com/discussions/general/352871

Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, 4, Article 972421. https://doi.org/10.3389/fnano.2022.972421

Santos, D., Saias, J., Quaresma, P., & Nogueira, V. B. (2021). Machine learning approaches to traffic accident analysis and hotspot prediction. *Computers*, *10*(12), 157. https://doi.org/10.3390/computers10120157

Senaviratna, N. S., & Cooray, T. M. J. A. (2019). Diagnosing Multicollinearity of Logistic Regression Model. *Asian Journal of Probability and Statistics*, 5(2), 1-9. https://doi.org/10.9734/ajpas/2019/v5i230132

Verdhan, V. (2020). *Supervised Learning with Python* (1st ed.) [E-book version]. Apress Berkeley, CA. https://doi-org.ezproxy-f.deakin.edu.au/10.1007/978-1-4842-6156-9

Victorian Government. (2023). *Red light cameras*. Victorian Government. https://www.vic.gov.au/red-light-cameras

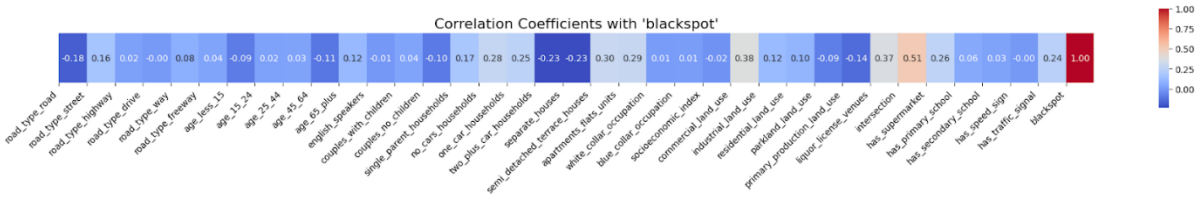
Visontay, E. (2023, May 14). Australia's road death toll jumps with fatalities still higher than pre-pandemic. *The Guardian*. https://www.theguardian.com/australia-news/2023/may/14/australia-road-death-toll-jumps-with-fatalities-still-higher-than-pre-pandemic

Zhang, L., Geisler, T., Ray, H., & Xie, Y. (2021). Improving logistic regression on the imbalanced data by a novel penalized log-likelihood function. *Journal of Applied Statistics*, 49(13), 3257–3277. https://doi.org/10.1080/02664763.2021.1939662

Appendices

Appendix A

Feature selection for first iteration in logistic regression model



In the first iteration of feature selection, the method employed was based on a threshold of the top 25% of absolute correlation coefficients in the heatmap. This approach aimed to identify features that demonstrated the strongest correlations with the target variable "blackspot". The following features were selected for this iteration:

1. 'no_cars_households'
2. 'one_car_households'
3. 'semi_detached_terrace_houses'
4. 'apartments_flats_units'
5. 'commercial_land_use'
6. 'liquor_license_venues'
7. 'intersection'
8. 'has_supermarket'
9. 'has_traffic_signal'

Appendix B

Feature selection for second iteration in logistic regression model

```
Optimization terminated successfully.
         Current function value: 0.222588
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:              blackspot   No. Observations:                 4248
Model:                          Logit   Df Residuals:                     4238
Method:                           MLE   Df Model:                            9
Date:                Thu, 10 Aug 2023   Pseudo R-squ.:                  0.3462
Time:                        23:56:57   Log-Likelihood:                -945.55
converged:                       True   LL-Null:                       -1446.1
Covariance Type:            nonrobust   LLR p-value:                9.623e-210
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           -3.6265      0.168    -21.530      0.000      -3.957      -3.296
no_cars_households              -0.2213      2.207     -0.100      0.920      -4.547       4.104
one_car_households               1.2729      0.647      1.968      0.049       0.006       2.540
semi_detached_terrace_houses     2.5555      0.859      2.975      0.003       0.872       4.239
apartments_flats_units           2.5888      1.103      2.346      0.019       0.426       4.751
commercial_land_use              2.2935      0.637      3.598      0.000       1.044       3.543
liquor_license_venues            0.1670      0.043      3.870      0.000       0.082       0.252
intersection                     2.6427      0.128     20.633      0.000       2.392       2.894
has_supermarket                  0.2258      0.548      0.412      0.680      -0.847       1.299
has_traffic_signal               0.5351      0.300      1.783      0.075      -0.053       1.123
================================================================================================
```

The second iteration of feature selection involved a more rigorous approach using backward elimination with p-value analysis using the statsmodels library. This method aimed to refine the feature set further by iteratively removing features that had the least statistical significance (higher p-value). As ‘no_cars_households’ had the highest p-value of 0.920, it was the first independent variable to be discarded, resulting in 8 features selected for this iteration:

1. 'one_car_households'
2. 'semi_detached_terrace_houses'
3. 'apartments_flats_units'
4. 'commercial_land_use'
5. 'liquor_license_venues'
6. 'Intersection'
7. ‘has_supermarket’
8. 'has_traffic_signal'

Through this process, the model underwent optimization by eliminating features with higher p-values, resulting in a more focused and relevant feature set for predicting blackspots.

Appendix C

Feature selection for third iteration in logistic regression model

```
Optimization terminated successfully.
         Current function value: 0.222589
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:              blackspot   No. Observations:                 4248
Model:                          Logit   Df Residuals:                     4239
Method:                           MLE   Df Model:                            8
Date:                Fri, 11 Aug 2023   Pseudo R-squ.:                  0.3462
Time:                        06:21:14   Log-Likelihood:                -945.56
converged:                       True   LL-Null:                       -1446.1
Covariance Type:            nonrobust   LLR p-value:                8.371e-211
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           -3.6242      0.167    -21.730      0.000      -3.951      -3.297
one_car_households               1.2461      0.589      2.115      0.034       0.091       2.401
semi_detached_terrace_houses     2.5183      0.775      3.250      0.001       1.000       4.037
apartments_flats_units           2.5599      1.066      2.401      0.016       0.470       4.650
commercial_land_use              2.2828      0.628      3.634      0.000       1.052       3.514
liquor_license_venues            0.1672      0.043      3.880      0.000       0.083       0.252
intersection                     2.6420      0.128     20.655      0.000       2.391       2.893
has_supermarket                  0.2224      0.546      0.407      0.684      -0.849       1.293
has_traffic_signal               0.5351      0.300      1.784      0.074      -0.053       1.123
================================================================================================
```

The third iteration of feature selection continued with the backward elimination and p-value analysis method using the statsmodels library. The goal was to further streamline the feature set by iteratively removing features with the least statistical significance. The subsequent attributes were chosen for this iteration following the exclusion of 'has_supermarket', which has a p-value equal to 0.684:

1. 'one_car_households'
2. 'semi_detached_terrace_houses'
3. 'apartments_flats_units'
4. 'commercial_land_use'
5. 'liquor_license_venues'
6. 'intersection'
7. 'has_traffic_signal'

By progressively removing features with higher p-values, the final feature set for the third iteration was refined to a subset that exhibited stronger statistical significance in predicting the likelihood of blackspots.